

A PTAS for Agnostically Learning Halfspaces

Amit Daniely *

June 26, 2015

Abstract

We present a PTAS for agnostically learning halfspaces w.r.t. the uniform distribution on the d dimensional sphere. Namely, we show that for every $\mu > 0$ there is an algorithm that runs in time $\text{poly}(d, \frac{1}{\epsilon})$, and is guaranteed to return a classifier with error at most $(1 + \mu)\text{opt} + \epsilon$, where opt is the error of the best halfspace classifier. This improves on Awasthi, Balcan and Long [2] who showed an algorithm with an (unspecified) constant approximation ratio. Our algorithm combines the classical technique of polynomial regression (e.g. [22, 16]), together with the new localization technique of [2].

*Department of Mathematics, Hebrew University, Jerusalem 91904, Israel. amit.daniely@mail.huji.ac.il

1 Introduction

In the problem of agnostically learning halfspaces, the learner is given an access to examples drawn from a distribution \mathcal{D} on $\mathbb{R}^d \times \{\pm 1\}$ and an accuracy parameter $\epsilon > 0$. It is required to output¹ a classifier $h : \mathbb{R}^d \rightarrow \{\pm 1\}$ whose error, $\text{Err}_{\mathcal{D}}(h) := \Pr_{(x,y) \sim \mathcal{D}}(h(x) \neq y)$, is at most² $\text{opt} + \epsilon$. Here, opt is the error of the best classifier of the form $h_w(x) = \text{sign}(\langle w, x \rangle)$. The learner is *efficient* if it runs in time $\text{poly}(d, \frac{1}{\epsilon})$. We note that we consider the general, *improper*, setting where the learner have the freedom to return a hypothesis that is not a halfspace classifier.

Halfspaces are extremely popular in practical applications, and have been extensively studied in Machine Learning, Statistics and Theoretical Computer Science (see section 1.2). Unfortunately, from a worst case perspective, the problem seems very hard: Best known efficient algorithms have a terrible approximation ratio of $\tilde{\Omega}(d)$. In the case of *proper learning*, where the output hypothesis must be a halfspace, agnostic learning is known to be \mathcal{NP} -hard. Even learning with a constant *approximation ratio*, where the returned classifier should have error $\leq \alpha \cdot \text{opt} + \epsilon$, is \mathcal{NP} -hard. In fact, even approximation ratio of $2^{\log^{0.99}(d)}$ is \mathcal{NP} -hard. In the general (improper) case, agnostic learning of halfspaces, and even agnostic learning with an approximation ratio of $2^{\log^{0.99}(d)}$, have been showed hard under various complexity assumptions (see section 1.2). In light of that, it is just natural to consider agnostic learning under various restrictions on the distribution \mathcal{D} . A very natural and widely studied such restriction [21, 20, 2, 16] is that the marginal distribution, $\mathcal{D}_{\mathbb{R}^d}$, is uniform on the sphere S^{d-1} .

Even under the uniform distribution, no efficient algorithms are known, and there is also an evidence that the problem is hard [19]. This lead researchers to consider *approximation algorithms*. The first approximation guarantee is due to [16], who showed an efficient regression based algorithm with approximation ratio of $\alpha = O\left(\sqrt{\log\left(\frac{1}{\text{opt}}\right)}\right)$. In an exciting recent work, [2] introduced a new algorithmic technique, called *localization*, and showed an efficient algorithm with an unspecified constant approximation ratio. In this paper, we advance this line of work further, and show a Polynomial Time Approximation Scheme (PTAS). Namely, we show:

Theorem 1.1 (main) *For every $\mu > 0$, there is an efficient algorithm for agnostically learning halfspaces under the uniform distribution with an approximation ratio of $(1 + \mu)$.*

As noted above, [19] showed that under a certain complexity assumption (hardness of learning sparse parity), there are no exact efficient algorithms (i.e., with approximation ratio $\alpha = 1$). In that case, our result is optimal.

Label Complexity: Our algorithm naturally fits to the *active learning* (e.g. [24]) setting. Often, a label is much more expensive than an example (e.g., when applying learning methods in biology, it might be the case that we have to make an experiment in order to get a label). It is therefore useful that algorithms will make economical use of labels. Our algorithm naturally have such property, as its *label complexity* (i.e., the number of labels it needs to see) is poly-logarithmic in $\frac{1}{\text{opt}}$ (see theorem 1.5 for a more detailed statement).

Interpolation between approximation and exact algorithms: A more precise statement of our result is that there exists an algorithm with runtime $\text{poly}\left(d^{\frac{\log^3(\frac{1}{\mu})}{\mu^2}}, \frac{1}{\epsilon}\right)$ that is guaranteed to return a classifier with error at most $(1 + \mu)\text{opt} + \epsilon$ for every $0 < \mu, \epsilon \leq 1$. Taking μ up to

¹Throughout, we require our algorithms to succeed with a constant probability (that can be standardly amplified by repetition).

²Note that opt might be > 0 , namely, we consider the “agnostic PAC learning” model [18].

$\frac{\epsilon}{2}$ and replacing ϵ with $\frac{\epsilon}{2}$, the error bound is $(1 + \frac{\epsilon}{2}) \text{opt} + \frac{\epsilon}{2} \leq \text{opt} + \epsilon$. Hence, we get an exact algorithm. The running time is $\text{poly}\left(d^{\frac{\log^3(\frac{1}{\epsilon})}{\epsilon^2}}\right)$, which almost matches the current state of the art – $\text{poly}\left(d^{\frac{1}{\epsilon^2}}\right)$ [16, 13].

Open questions: Obvious open questions are to extend our results to more distributions (uniform on $\{\pm 1\}^d$, permutation-invariant, product, log-concave, ...) and more problems (learning intersection of halfspaces, functions of halfspaces, ...). In addition, as opposed to previous approximation algorithms [2, 16], our algorithm does not always return a halfspace classifier. A natural open question is therefore to find a *proper* PTAS.

1.1 Algorithmic Components, The PTAS, and Proof Outline

Our algorithm and its analysis build on and combine various algorithmic and proof techniques that were previously used for learning halfspaces. This includes regression based algorithms (e.g. [25, 16]), polynomial approximations of the sign function (e.g. [25, 16, 12, 13]) and localization techniques [2]. In this section we outline these techniques and the way we use them. Then, we present our PTAS, state its properties (theorem 1.5), and describe the course of the proof. The full proof is in sections 2 and 3.

1.1.1 Some preliminaries

Noise tolerance is a measure to evaluate the performance of learning algorithms, that is essentially equivalent to the approximation ratio. Yet, we find it slightly more convenient for the technical exposition. We say that a learning algorithm tolerates noise rate of $0 < f(\eta) < \eta$ (w.r.t. halfspaces) if, when running on input $0 < \eta < 1$, it is guaranteed to return a hypothesis with error $\leq \eta$, provided that $\text{opt} \leq f(\eta)$. We say that such an algorithm is *efficient* if it runs in time $\text{poly}\left(d, \frac{1}{\eta}\right)$. We note that given a learning algorithm that tolerates noise rate of $\frac{\eta}{\alpha}$, for some $\alpha > 1$, it is not hard to construct an algorithm with approximation ratio of α , and the running time grows only by a factor of $\text{poly}\left(\frac{1}{\epsilon}\right)$: Indeed, in order to return a hypothesis with error $\leq \alpha \cdot \text{opt} + \epsilon$, we can run the algorithm with $\alpha \cdot \text{opt} \leq \eta \leq \alpha \cdot \text{opt} + \epsilon$. We can find such an η by trying $\eta = k\epsilon$ for $k = 1, 2, \dots, \lceil \frac{1}{\epsilon} \rceil$. **Notation.** Let \mathcal{D} be a distribution on a space X . For $Y \subset X$ we denote by $\mathcal{D}|_Y$ the restriction of \mathcal{D} to Y . If \mathcal{D} is a distribution on $X \times \{\pm 1\}$ we denote by \mathcal{D}_X the marginal distribution on X . If \mathcal{D} is a distribution on S^{d-1} (resp. $S^{d-1} \times \{\pm 1\}$) and $w \in S^{d-1}$, we define the *projection of \mathcal{D} on w* as follows: If $x \sim \mathcal{D}$ (resp. $(x, y) \sim \mathcal{D}$) then \mathcal{D}_w is the distribution (on $[-1, 1]$) of the random variable $\langle w, x \rangle$. For a distribution \mathcal{D} on a space X and a function $f : S^{d-1} \rightarrow \mathbb{R}$, we denote $\|f\|_{p, \mathcal{D}} = (\mathbb{E}_{x \sim \mathcal{D}} |f(x)|^p)^{\frac{1}{p}}$. We will sometimes abuse notation and use $\|f\|_{p, \mathcal{D}}$ instead of $\|f\|_{p, \mathcal{D}_{S^{d-1}}}$ even when \mathcal{D} is a distribution on $S^{d-1} \times \{\pm 1\}$. We denote by $\theta(w, w^*) = \cos^{-1}(\langle w, w^* \rangle)$ the angle between two vectors $w, w^* \in S^{d-1}$. We will frequently use the fact that for uniform $x \in S^{d-1}$ we have $\Pr(h_{w^*}(x) \neq h_w(x)) = \frac{\theta(w, w^*)}{\pi}$. We denote by $\text{POL}_{r, d}$ the space of d -variate polynomials of degree $\leq r$. For $w \in S^{d-1}$ and $\gamma > 0$ we let $T_{d, \gamma}(w) := \{u \in S^{d-1} : |\langle w, u \rangle| \leq \gamma\}$.

1.1.2 Polynomial ℓ_1 -regression for classification

The output of a classification algorithm is a (description of a) hypothesis $h : S^{d-1} \rightarrow \{\pm 1\}$. Often, the returned hypothesis is of the form $h(x) = \text{sign}(f(x))$, for some real valued function $f : S^{d-1} \rightarrow \mathbb{R}$. To conveniently dealing with such hypotheses, we introduce some terminology. We denote the standard (zero-one) loss of f by $\text{Err}_{\mathcal{D}}(f) = \text{Err}_{\mathcal{D}}(\text{sign}(f))$. We also consider the ℓ_1 -loss,

$\text{Err}_{\mathcal{D},1}(f) = \mathbb{E}_{(x,y) \sim \mathcal{D}} |f(x) - y|$. We note that for $f : S^{d-1} \rightarrow \mathbb{R}$, since $\frac{|\text{sign}(z)-1|}{2} \leq |z - 1|$ for all z , we have

$$\begin{aligned} \text{Err}_{\mathcal{D}}(f) &= \mathbb{E}_{(x,y) \sim \mathcal{D}} \frac{|\text{sign}(yf(x)) - 1|}{2} \\ &\leq \mathbb{E}_{(x,y) \sim \mathcal{D}} |yf(x) - 1| \\ &= \mathbb{E}_{(x,y) \sim \mathcal{D}} |f(x) - y| = \text{Err}_{\mathcal{D},1}(f) \end{aligned}$$

Thus, by finding f with small ℓ_1 -error we can find a good classifier. The motivation for moving from the 0-1 loss to the ℓ_1 loss is the convexity of the ℓ_1 loss, which enables the use of convex optimization. Concretely, for “nice enough” convex set, \mathcal{F} , of functions from S^{d-1} to \mathbb{R} , it is possible to efficiently find (both in terms of number of examples and time) $f \in \mathcal{F}$ with ℓ_1 error almost as small as $\min_{f \in \mathcal{F}} \text{Err}_{\mathcal{D},1}(f)$. Now, for a classifier $h : S^{d-1} \rightarrow \{\pm 1\}$ we have

$$\begin{aligned} \text{Err}_{\mathcal{D}}(f) \leq \text{Err}_{\mathcal{D},1}(f) &= \mathbb{E}_{(x,y) \sim \mathcal{D}} |f(x) - y| \\ &\leq \mathbb{E}_{(x,y) \sim \mathcal{D}} |f(x) - h(x)| + \mathbb{E}_{(x,y) \sim \mathcal{D}} |h(x) - y| \\ &= \|f - h\|_{1,\mathcal{D}} + 2 \text{Err}(h) \end{aligned} \tag{1}$$

Thus, if we minimize the ℓ_1 -loss over a collection of functions that is large enough to contain a good ℓ_1 -approximation of the best halfspace classifier, we can find a function whose ℓ_1 -error, and therefore also the 0-1 error, is almost as good as the 0-1 error of the best halfspace classifier. Methods that follow the above spirit have been extensively studied in computational learning theory. Concretely, [16] suggested the following algorithm: First, find $P \in \text{POL}_{r,d}$ that minimizes the empirical ℓ_1 -error on the given sample³. Then, find a classifier that makes the least number of errors on the given sample, among all classifiers of the form $x \mapsto \text{sign}(P(x) - a)$ for $a \in \mathbb{R}$. We note that the second step is required in order to overcome the factor of 2 in equation (1). They used that algorithm to show:

Theorem 1.2 [16] *There is an algorithm with runtime $\text{poly}(d^r, \frac{1}{\epsilon})$ such that, for every distribution \mathcal{D} on $S^{d-1} \times \{\pm 1\}$ and every $h : S^{d-1} \rightarrow \{\pm 1\}$, it returns $P \in \text{POL}_{r,d}$ with $\text{Err}_{\mathcal{D}}(P) \leq \text{Err}_{\mathcal{D}}(h) + \min_{P' \in \text{POL}_{r,d}} \|h - P'\|_{1,\mathcal{D}} + \epsilon$.*

1.1.3 Learning halfspaces using sign approximations

To use theorem 1.2 for learning halfspaces, we need to prove the existence of low degree polynomials P such that $\|h - P\|_{1,\mathcal{D}}$ is small, where h is a halfspace classifier. As explained below, this is naturally done by approximating the *sign function*, $\text{sign}(x) = \begin{cases} 1 & x > 0 \\ -1 & x \leq 0 \end{cases}$, with respect to an appropriate proximity measure.

Suppose that $w^* \in S^{d-1}$ defines the optimal halfspace and let \mathcal{D}_{w^*} be the projection of \mathcal{D} on w^* . For a univariate polynomial $p \in \text{POL}_{r,1}$, consider the d -variate polynomial $P \in \text{POL}_{r,d}$ given by $P(x) = p(\langle w^*, x \rangle)$. We have

$$\begin{aligned} \|P - h_{w^*}\|_{1,\mathcal{D}} &= \mathbb{E}_{x \sim \mathcal{D}_{S^{d-1}}} [|p(\langle w^*, x \rangle) - \text{sign}(\langle w^*, x \rangle)|] \\ &= \mathbb{E}_{x \sim \mathcal{D}_{w^*}} [|p(x) - \text{sign}(x)|] \\ &= \|p - \text{sign}\|_{1,\mathcal{D}_{w^*}} \end{aligned} \tag{2}$$

³I.e., if the sample is $(x_1, y_1), \dots, (x_m, y_m) \in S^{d-1} \times \{\pm 1\}$, find $P \in \text{POL}_{r,d}$ that minimizes $\frac{1}{m} \sum_{i=1}^m |P(x_i) - y_i|$.

Therefore, in order to find a good ℓ_1 approximation for h_{w^*} w.r.t. \mathcal{D} , we can find a good ℓ_1 approximation for sign w.r.t. \mathcal{D}_{w^*} .

Approximating the sign function is a central component in many papers about halfspaces [6, 12, 12, 16, 25]. These papers needed to find approximation of the sign function w.r.t. relatively well studied proximity measures, such as the ℓ_∞ norm, or the ℓ_1 and ℓ_2 norms w.r.t. the Gaussian distribution. Therefore, some of these papers used basis expansion methods (Fourier, Hermite, Chebyshev, ...). In this paper we need to find ℓ_1 approximation w.r.t. messier distributions. Therefore, we use a somewhat more flexible approach, similar to the one used in [12]. We rely on techniques from approximation theory [11]. In particular, our main tool for constructing polynomials will be the celebrated Jackson's theorem.

Theorem 1.3 (Jackson, [11]) *For every L -lipschitz function $f : [-1, 1] \rightarrow \mathbb{R}$ and $r \in \mathbb{N}$ there is a degree r polynomial p such that $\|p - f\|_{\infty, [-1, 1]} \leq \frac{6L}{r}$*

1.1.4 Localization

An additional algorithmic component we will use, except polynomial regression, is *localization in the instance and the hypotheses space* (e.g. [3, 2]). The basic idea is the following. Suppose that $w^* \in S^{d-1}$ defines the optimal halfspace. Suppose furthermore that we have found (say, using some simple algorithm) a vector $w \in S^{d-1}$ that defines a halfspace with a relatively small error. The facts that the marginal distribution is uniform and $\text{Err}(h_w)$ is small have two relevant consequences:

- We know that the optimal vector, w^* , is close to w .
- Hence, if $|\langle w, x \rangle|$ is large, then $h_{w^*}(x) = h_w(x)$ and therefore we know $h_{w^*}(x)$.

These two properties enable us to “localize the learning” and concentrate only on hypotheses $h_{w'}$ with w' close to w , and on instances x with small $|\langle w, x \rangle|$. We will use this idea directly in our algorithm. In addition, we will use, as a black-box, the localization-based algorithm of [2]. Their algorithm starts with a crude approximation $w_1 \in S^{d-1}$ of the optimal halfspace w^* . Then, it finds w_2 that minimizes the *hinge loss* $\mathbb{E}_{\mathcal{D}|_{T \times \{\pm 1\}}}(1 - \langle w, yx \rangle)_+$ on the restriction of \mathcal{D} to some small strip $T = \{x \in S^{d-1} \mid |\langle w, x \rangle| \leq \gamma\}$. Then, it continue in this manner to find better and better w_i 's. Awasthi, Balcan and Long used their algorithm to show:

Theorem 1.4 [2] *There is an efficient learning algorithm with label complexity $\text{poly}\left(d, \log\left(\frac{1}{\eta}\right)\right)$ that tolerates noise rate of $\frac{\eta}{\alpha_0}$ for some universal constant $\alpha_0 > 1$. Moreover, the algorithm is proper, that is, its output is a halfspace.*

1.1.5 The PTAS and its analysis

In a nutshell, our algorithm first find (step 1) a “rough estimation”, w , of w^* . Then, it “localizes the learning” and apply more computation power (step 3), to a small strip T that is closed to h_w 's decision boundary, and therefore, intuitively, we are less certain about h_w 's prediction.

Algorithm 1 A PTAS for agnostically learning halfspaces w.r.t. the uniform distribution

Input: $0 < \eta \leq 1$ and access to samples from a distribution \mathcal{D} on $S^{d-1} \times \{\pm 1\}$.

Parameters: $r \in \mathbb{N}$, $\beta > 0$ and $\gamma > 0$.

- 1: Find, using [2] (theorem 1.4), a vector $w \in S^{d-1}$ with $\text{Err}_{\mathcal{D}}(h_w) \leq \alpha_0 \eta$
- 2: Let $T = T_{d,\gamma}(w) := \{u \in S^{d-1} : |\langle w, u \rangle| \leq \gamma\}$.
- 3: Find, using [16] (theorem 1.2), $P \in \mathbf{POL}_{r,d}$ with

$$\text{Err}_{\mathcal{D}|_T}(P) \leq \text{Err}_{\mathcal{D}|_T}(h_{w^*}) + \min_{P' \in \mathbf{POL}_{r,d}} \|h_{w^*} - P'\|_{1,\mathcal{D}|_T} + \beta$$

where h_{w^*} is an optimal halfspace classifier w.r.t. \mathcal{D} .

- 4: With probability $\frac{1}{2}$ return h_w , and w.p. $\frac{1}{2}$ return the classifier

$$h(x) = \begin{cases} h_w(x) & |\langle w, x \rangle| > \gamma \\ \text{sign}(P(x)) & |\langle w, x \rangle| \leq \gamma \end{cases}$$

Theorem 1.5 (main – detailed) *With appropriate choice of the parameters r, β, γ (depending on $0 < \mu, \eta \leq 1$), algorithm 1 satisfies:*

- It tolerates noise rate of $(1 - \mu)\eta$.
- It runs in time $\text{poly}\left(d^{\frac{\log^3(\frac{1}{\mu})}{\mu^2}}, \frac{1}{\eta}\right)$.
- Its label complexity is $\text{poly}\left(d^{\frac{\log^3(\frac{1}{\mu})}{\mu^2}}, \log\left(\frac{1}{\eta}\right)\right)$.

Proof outline. To prove theorem 1.5, we must show that we can choose the parameters so that the time and label complexity are as stated, and under the assumption that $\text{Err}_{\mathcal{D}}(h_{w^*}) \leq (1 - \mu)\eta$, the error of the returned classifier satisfies $\text{Err}_{\mathcal{D}}(h) \leq \eta$. Below, we explain how we do that. We would naturally like to decompose the error into two parts:

$$\begin{aligned} \text{Err}_{\mathcal{D}}(h) &= \Pr_{(x,y) \sim \mathcal{D}}(x \notin T) \cdot \text{Err}_{\mathcal{D}|_{T^c \times \{\pm 1\}}}(h) + \Pr_{(x,y) \sim \mathcal{D}}(x \in T) \cdot \text{Err}_{\mathcal{D}|_{T \times \{\pm 1\}}}(h) \\ &= \Pr_{(x,y) \sim \mathcal{D}}(x \notin T) \cdot \text{Err}_{\mathcal{D}|_{T^c \times \{\pm 1\}}}(h_w) + \Pr_{(x,y) \sim \mathcal{D}}(x \in T) \cdot \text{Err}_{\mathcal{D}|_{T \times \{\pm 1\}}}(P) \end{aligned} \quad (3)$$

We first handle the former summand using a localization lemma (lemma 2.1 below). We show that for $\gamma = \Theta\left(\frac{\eta \sqrt{\log(\frac{1}{\mu})}}{\sqrt{d}}\right)$, the probability that $h_w(x) \neq h_{w^*}(x)$ outside the strip T , is $\leq \frac{\mu\eta}{2}$. Hence, on the complement of T , the returned classifier, that coincides with h_w , is as good as h_{w^*} , up to an additive error of $\frac{\mu\eta}{2}$. Concretely,

$$\Pr_{(x,y) \sim \mathcal{D}}(x \notin T) \cdot \text{Err}_{\mathcal{D}|_{T^c \times \{\pm 1\}}}(h_w) \leq \Pr_{(x,y) \sim \mathcal{D}}(x \notin T) \cdot \text{Err}_{\mathcal{D}|_{T^c \times \{\pm 1\}}}(h_{w^*}) + \frac{\mu\eta}{2}. \quad (4)$$

It remains to handle the latter summand in equation (4). It is enough to show that

$$\Pr_{(x,y) \sim T}(x \in T) \cdot \text{Err}_{\mathcal{D}|_{T \times \{\pm 1\}}}(P) \leq \Pr_{(x,y) \sim T}(x \in T) \cdot \text{Err}_{\mathcal{D}|_{T \times \{\pm 1\}}}(h_{w^*}) + \frac{\mu\eta}{2} \quad (5)$$

Indeed, in that case it follows from equations (3), (4) and (5) that

$$\begin{aligned} \text{Err}_{\mathcal{D}}(h) &\leq \Pr_{(x,y) \sim \mathcal{D}}(x \notin T) \cdot \text{Err}_{\mathcal{D}|_{T^c \times \{\pm 1\}}}(h_{w^*}) + \Pr_{(x,y) \sim T}(x \in T) \cdot \text{Err}_{\mathcal{D}|_{T \times \{\pm 1\}}}(h_{w^*}) + \mu\eta \\ &= \text{Err}_{\mathcal{D}}(h_{w^*}) + \mu\eta \leq (1 - \mu)\eta + \mu\eta = \eta. \end{aligned}$$

To prove equation (5) we first note that $\Pr_{(x,y) \sim \mathcal{D}}(x \in T) = \Theta\left(\eta \sqrt{\log\left(\frac{1}{\mu}\right)}\right)$. Hence, it is enough to show that for suitable choice of r and β , $\text{Err}_{\mathcal{D}|_{T \times \{\pm 1\}}}(P) \leq \text{Err}_{\mathcal{D}|_{T \times \{\pm 1\}}}(h_{w^*}) + \frac{\mu}{C \sqrt{\log\left(\frac{1}{\mu}\right)}}$ for large enough constant $C > 0$. By theorem 1.2, it is enough to choose $\beta = \frac{\mu}{2C \sqrt{\log\left(\frac{1}{\mu}\right)}}$, and large enough r so that $\min_{P' \in \text{POL}_{r,d}} \|h - P'\|_{1, \mathcal{D}|_{T \times \{\pm 1\}}} \leq \frac{\mu}{2C \sqrt{\log\left(\frac{1}{\mu}\right)}}$.

As we show, $r = O\left(\frac{\log^3\left(\frac{1}{\mu}\right)}{\mu^2}\right)$ suffices. To do that, by equation (2), it is enough to find a polynomial of degree $O\left(\frac{\log^3\left(\frac{1}{\mu}\right)}{\mu^2}\right)$ that approximates the sign function up to an ℓ_1 -error of $\frac{\mu}{2C \sqrt{\log\left(\frac{1}{\mu}\right)}}$ w.r.t the distribution $(\mathcal{D}|_{T \times \{\pm 1\}})_{w^*}$. This is done in section 3, in three steps:

1. We first (section 3.1) show how to find polynomials that approximate the sign function on all the points of a given segment $[-a, a]$, except the area that is very close to the origin, say $[-\epsilon, \epsilon]$. To this end, we invoke Jackson's theorem (theorem 1.3) to find a polynomial that roughly (up to an error of, say, 0.1) approximates the sign function on the mentioned regime. Namely, we find a polynomial p of degree $O\left(\frac{a}{\epsilon}\right)$ that maps $[-a, -\epsilon]$ (resp. $[\epsilon, a]$) to $[-1.1, -0.9]$ (resp. $[0.9, 1.1]$). To move from accuracy of 0.1 to accuracy of some small $\tau > 0$, we compose p with another polynomial r that maps $[-1.1, -0.9]$ (resp. $[0.9, 1.1]$) to $[-1 - \tau, -1 + \tau]$ (resp. $[1 - \tau, 1 + \tau]$). Using the Taylor expansion of the *error function* $\text{erf}(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$, we show that there exists such r of degree $O\left(\log\left(\frac{1}{\tau}\right)\right)$.
2. In the second step (section 3.2), we will find ℓ_1 approximations for distributions with strong tail bounds (namely, distributions that have density function bounded by $2 \exp\left(-\frac{x^2}{32}\right)$ on a certain domain). We use step 1 to find polynomials that approximate the sign function in ℓ_∞ on a relatively large area, and use the tail bounds and lemma 3.4 to neglect the ℓ_1 norm on the complement of that area.
3. In the last step (section 3.3), using basic facts about high dimensional spherical geometry, we show that the distribution $(\mathcal{D}|_{T \times \{\pm 1\}})_{w^*}$ have strong enough tail bounds.

1.2 Related work

Upper bounds. Statistical aspects of learning halfspaces have been extensively studied (e.g. [26]). Halfspaces are efficiently learnable in the *realizable case*, when $\text{opt} = 0$. This is done using the ERM algorithm [26] that efficiently find, using linear programming, a halfspace that makes no errors on the given sample. For agnostic, distribution free learning, the best known efficient algorithm [17] have an approximation ratio of $O(d)$, and the best known exact algorithm is the naive (exponential time) algorithm that go over all halfspaces and return the one with minimal error on the

given sample. Under distributional assumptions, better algorithms are known. Under the uniform distribution, [16] and [2] presented efficient algorithms with approximation ratios $\sqrt{\log\left(\frac{1}{\text{opt}}\right)}$ and $O(1)$ respectively. The best known exact algorithm [16] runs in time $d^{O\left(\frac{1}{\epsilon^2}\right)}$ (as follows from [13]). For log-concave distributions, [21] and [2] presented efficient algorithms with approximation ratios $O\left(\frac{\log\left(\frac{1}{\text{opt}}\right)}{\text{opt}^2}\right)$ and $O\left(\log^2\left(\frac{1}{\text{opt}}\right)\right)$ respectively. The best known exact algorithm [16] runs in time $d^{f(\epsilon)}$. In learning halfspaces with margin⁴ $\gamma > 0$, best known algorithms [23, 6] have approximation ratio of $\frac{1/\gamma}{\log(1/\gamma)}$, while the best known exact algorithm [25] runs in time $\left(\frac{1}{\epsilon}\right)^{O\left(\frac{\log(1/\gamma)}{\gamma}\right)}$.

Lower bounds. Hardness of (distribution free) agnostic learning of halfspaces is known to follow from several complexity assumptions including hardness of learning parity [16] (this result even rules out learning under the uniform distribution on $\{\pm 1\}^d$), hardness of the shortest vector problem [14], and hardness of refuting random K -SAT formulas [8]. Hardness of learning sparse parity implies hardness of agnostic learning under the uniform distribution on S^{d-1} [19]. For every $\tau > 0$, hardness of agnostic learning of halfspaces with an approximation ratio of $2^{\log^{1-\tau}(d)}$ follows from hardness of refuting random K -XOR formulas [7] (see also [9]). For *proper* learning of halfspaces, super constant $(2^{\log^{1-\tau}(d)})$ for every $\tau > 0$ lower bounds on the best approximation ratio are known, assuming $\mathcal{NP} \neq \mathcal{RP}$ [1, 15, 14]. Finally, lower bounds on concrete families of algorithms were studied in [4, 10]

2 Proof of theorem 1.5

For localization arguments, we will use the following lemma.

Lemma 2.1 (localization) *Let $w, w^* \in S^{n-1}$ and let \mathcal{D} be a distribution of $S^{d-1} \times \{\pm 1\}$ such that $\mathcal{D}|_{S^{d-1}}$ is uniform.*

- We have $\frac{\theta(w, w^*)}{\pi} \leq \text{Err}_{\mathcal{D}}(w) + \text{Err}_{\mathcal{D}}(w^*)$.
- If $x \in S^{d-1}$ is a uniform vector, then for every $r > 0$,

$$\Pr(h_w(x) \neq h_{w^*}(x) \text{ and } |\langle x, w \rangle| > r \cdot \theta(w, w^*)) \leq \frac{4 \cdot \theta(w, w^*)}{\pi} \exp\left(-\frac{1}{8}r^2d\right)$$

Proof For the first part we note that $\Pr_{x \sim \mathcal{D}}(h_w(x) \neq h_{w^*}(x)) = \frac{\theta(w, w^*)}{\pi}$, while on the other hand,

$$\Pr_{x \sim \mathcal{D}}(h_w(x) \neq h_{w^*}(x)) \leq \Pr_{(x, y) \sim \mathcal{D}}(h_w(x) \neq y) + \Pr_{(x, y) \sim \mathcal{D}}(h_{w^*}(x) \neq y) .$$

For the second part, let $V \subset \mathbb{R}^d$ be the 2-dimensional space spanned by w, w^* , let $P_V : \mathbb{R}^d \rightarrow V$ be the orthogonal projection V , and let $B \subset V$ be the ball of radius r around 0. We have

$$|\langle w^*, x \rangle - \langle w, x \rangle| = |\langle w^* - w, P_V(x) \rangle| \leq \|w - w^*\| \cdot \|P_V(x)\| \leq \theta(w, w^*) \cdot \|P_V(x)\| .$$

⁴In this problem the distribution is supported in the unit ball, and the algorithm should compete with all classifiers that predict like a halfspace classifier h_w , except that they give no prediction (and therefore err) for instances that are within distance γ of the decision boundary of h_w .

Therefore, if $P_V(x) \in B$ and $|\langle x, w \rangle| > r \cdot \theta(w, w^*)$ then $h_w(x) = h_{w^*}(x)$. It follows that

$$\begin{aligned} \Pr(h_w(x) \neq h_{w^*}(x) \text{ and } |\langle x, w \rangle| > r \cdot \theta(w, w^*)) &= \Pr(h_w(x) \neq h_{w^*}(x) \mid P_V(x) \notin B) \cdot \Pr(P_V(x) \notin B) \\ &= \frac{\theta(w, w^*)}{\pi} \cdot \Pr(P_V(x) \notin B). \end{aligned}$$

Finally, let $e_1, e_2 \in V$ be an orthonormal basis. Note that if $|\langle x, e_1 \rangle| \leq \frac{r}{\sqrt{2}}$ and $|\langle x, e_2 \rangle| \leq \frac{r}{\sqrt{2}}$ then $P_V(x) \in B$. Hence, we have

$$\Pr(P_V(x) \notin B) \leq \Pr\left(|\langle x, e_1 \rangle| > \frac{r}{\sqrt{2}}\right) + \Pr\left(|\langle x, e_2 \rangle| > \frac{r}{\sqrt{2}}\right) \leq 4 \exp\left(-\frac{1}{8}r^2d\right).$$

Here, the last inequality follows from the well known measure concentration bound according which for every $e \in S^{d-1}$ and $\sigma > 0$ we have $\Pr(|\langle x, e \rangle| \geq \sigma) \leq 2 \exp(-\frac{1}{4}\sigma^2d)$. \square

To approximate h_{w^*} , we will need to find low degree ℓ_1 approximation of h_{w^*} w.r.t. the distribution $\mathcal{D}|_T$. Such approximations are given in the following two lemmas. The first lemma is from [12] (see a proof in section 3. For a stronger version, with $r = O(\frac{1}{\tau^2})$, see [13]). The proof of the second lemma is established by approximating the sign function (as explained in section 1.1.3) and is given in section 3.

Lemma 2.2 (uniform halfspaces approximation, [12]) *Let \mathcal{D} be the uniform distribution on S^{d-1} and let $w^* \in S^{d-1}$. For every $\tau > 0$ there is $P \in \text{POL}_{r,d}$, for $r = O\left(\frac{\log^2(1/\tau)}{\tau^2}\right)$ such that $\|h_{w^*} - P\|_{1,\mathcal{D}} < \tau$.*

Lemma 2.3 (halfspaces approximation on a strip) *Let w, w^* be two vectors with $\theta = \theta(w, w^*)$ and let $\frac{1}{2} > \gamma > 0$. Let \mathcal{D} be the distribution on S^{d-1} that is the restriction of the uniform distribution to $T_{d,\gamma}(w)$. Then, for every $0 < \tau < \frac{\sin(\theta)}{2\gamma\sqrt{d}}$ there is $P \in \text{POL}_{r,d}$, for $r = O\left(\frac{\log^2(1/\tau)}{\tau^2}\right)$ such that $\|h_{w^*} - P\|_{1,\mathcal{D}} < \tau$.*

Lastly, we will also rely on the following complexity analysis of algorithm 1.

Lemma 2.4 (complexity analysis) *The runtime of algorithm 1 is $\text{poly}\left(d^r, \frac{1}{\beta}, \frac{1}{\gamma}, \frac{1}{\eta}\right)$ and the label complexity is $\text{poly}\left(d^r, \frac{1}{\eta}, \log\left(\frac{1}{\eta}\right)\right)$.*

Proof The runtime of step 1 is $\text{poly}\left(d, \frac{1}{\eta}\right)$, while the label complexity is $\text{poly}\left(d, \log\left(\frac{1}{\eta}\right)\right)$. For step 3, we can apply the [16] algorithm on $\text{poly}\left(d^r, \frac{1}{\eta}\right)$ examples and labels from the distribution $\mathcal{D}|_T$. We can get these many examples by sampling $\text{poly}\left(d^r, \frac{1}{\beta}, \frac{1}{\Pr_{\mathcal{D}}(T \times \{\pm 1\})}\right)$ examples from \mathcal{D} and keep and expose the labels of only the first $\text{poly}\left(d^r, \frac{1}{\beta}\right)$ examples that fell in T . It is not hard to see that $\Pr_{\mathcal{D}}(T \times \{\pm 1\}) \geq \Omega\left(\min\left(\gamma\sqrt{d}, 1\right)\right)$. Hence, the runtime of step 3 is $\text{poly}\left(d^r, \frac{1}{\beta}, \frac{1}{\gamma}\right)$. To summarize, the total runtime is $\text{poly}\left(d^r, \frac{1}{\beta}, \frac{1}{\gamma}, \frac{1}{\eta}\right)$ and the label complexity is $\text{poly}\left(d^r, \frac{1}{\beta}, \log\left(\frac{1}{\eta}\right)\right)$. \square

We are now ready to prove theorem 1.5.

Proof (of theorem 1.5) We will first deal with the case that $\eta > \frac{1}{2(1+\alpha_0)}$. In that case we won't use localization, that is we will choose $\gamma = 1$ (in that case our algorithm is essentially the algorithm of [16]). We will choose $\beta = \frac{\mu\eta}{2}$, and $r = O\left(\frac{\log^2(1/(\mu\eta))}{(\mu\eta)^2}\right) = O\left(\frac{\log^2(1/\mu)}{\mu^2}\right)$ that is large enough so that $\min_{P' \in \text{POL}_{r,d}} \|h_{w^*} - P'\|_{1,\mathcal{D}|_T} \leq \frac{\mu\eta}{2}$ (this is possible according to lemma 2.2). In that case, the algorithm will, w.p. $\frac{1}{2}$, return the hypothesis $\text{sign}(P)$ for the polynomial P that was found in step 3. We have

$$\text{Err}_{\mathcal{D}}(P) \leq \text{Err}_{\mathcal{D}}(h_{w^*}) + \frac{\mu\eta}{2} + \frac{\mu\eta}{2}.$$

By assumption, $\text{Err}_{\mathcal{D}}(h_{w^*}) \leq (1 - \mu)\eta$. Hence, $\text{Err}_{\mathcal{D}}(P) \leq \eta$, as required. It also follows from lemma 2.4 that the runtime and label complexity are $\text{poly}\left(d^{\frac{\log^2(1/\mu)}{\mu^2}}\right)$ (note that η is bounded from below by a constant) as stated.

Next, we deal with the case that $\eta \leq \frac{1}{2(1+\alpha_0)}$. We will show that it is possible to choose $r = \Theta\left(\frac{\log^3(\frac{1}{\mu})}{\mu^2}\right)$, $\beta = \theta\left(\frac{\mu}{\sqrt{\log(\frac{1}{\mu})}}\right)$ and $\gamma = \Theta\left(\frac{\eta\sqrt{\log(\frac{1}{\mu})}}{\sqrt{d}}\right)$ for which the algorithm will have the desired properties. Also, by lemma 2.4, for such a choice of parameters, the runtime and label complexity are as stated.

Let w^* be the vector defining the optimal halfspace. By assumption, $\text{Err}_{\mathcal{D}}(h_{w^*}) \leq (1 - \mu)\eta$. Let w be the vector found in step 1, and let P be the polynomial found in step 3. We first claim that we can assume w.l.o.g. that

$$\frac{\theta}{\pi} := \frac{\theta(w, w^*)}{\pi} \geq \mu\eta. \quad (6)$$

Indeed, otherwise, we will have

$$\begin{aligned} \text{Err}_{\mathcal{D}}(h_w) &\leq \text{Err}_{\mathcal{D}}(h_{w^*}) + \Pr_{(x,y) \sim \mathcal{D}}(h_w(x) \neq h_{w^*}(x)) \\ &= \text{Err}_{\mathcal{D}}(h_{w^*}) + \frac{\theta}{\pi} \leq (1 - \mu)\eta + \mu\eta < \eta \end{aligned}$$

and in that case the algorithm will return, in the last step, w.p. $\frac{1}{2}$, a hypothesis with error at most η , as required.

Let $h(x) = \begin{cases} h_w(x) & |\langle w, x \rangle| > \gamma \\ \text{sign}(P(x)) & |\langle w, x \rangle| \leq \gamma \end{cases}$. It is enough to show that $\text{Err}_{\mathcal{D}}(h) \leq \eta$. Let $T = T_{d,\gamma}(w) := \{u \in S^{d-1} : |\langle w, u \rangle| \leq \gamma\}$. The error of h is

$$\begin{aligned} \text{Err}_{\mathcal{D}}(h) &= \Pr_{(x,y) \sim \mathcal{D}}(h_w(x) \neq y \text{ and } |\langle w, x \rangle| > \gamma) + \Pr_{(x,y) \sim \mathcal{D}}(\text{sign}(P(x)) \neq y \text{ and } |\langle w, x \rangle| \leq \gamma) \\ &\leq \Pr_{(x,y) \sim \mathcal{D}}(h_w(x) \neq h_{w^*}(x) \text{ and } |\langle w, x \rangle| > \gamma) + \Pr_{(x,y) \sim \mathcal{D}}(h_{w^*}(x) \neq y \text{ and } |\langle w, x \rangle| > \gamma) \\ &\quad + \Pr_{(x,y) \sim \mathcal{D}}(x \in T) \cdot \text{Err}_{\mathcal{D}|_T}(P) \end{aligned} \quad (7)$$

By the first part of lemma 2.1 we have

$$\frac{\theta}{\pi} \leq \text{Err}_{\mathcal{D}}(h_w) + \text{Err}_{\mathcal{D}}(h_{w^*}) \leq (1 + \alpha_0)\eta. \quad (8)$$

By the second part of lemma 2.1 we have

$$\begin{aligned} \Pr_{(x,y) \sim \mathcal{D}} (h_w(x) \neq h_{w^*}(x) \text{ and } |\langle w, x \rangle| > \gamma) &\leq 4(1 + \alpha_0)\eta \exp\left(-\frac{1}{8} \left(\frac{\gamma}{\theta}\right)^2 d\right) \\ &\leq 4(1 + \alpha_0)\eta \exp\left(-\frac{1}{8} \left(\frac{\gamma}{(1 + \alpha_0)\pi\eta}\right)^2 d\right) \end{aligned}$$

Now, by an appropriate choice of $\gamma = \Theta\left(\frac{\eta\sqrt{\log(\frac{1}{\mu})}}{\sqrt{d}}\right)$, we get

$$\Pr_{(x,y) \sim \mathcal{D}} (h_w(x) \neq h_{w^*}(x) \text{ and } |\langle w, x \rangle| > \gamma) \leq \frac{\mu\eta}{2}. \quad (9)$$

We next deal with the term $\Pr_{(x,y) \sim \mathcal{D}} (x \in T) \cdot \text{Err}_{\mathcal{D}|_T}(P)$. Since $\gamma = \Theta\left(\frac{\eta\sqrt{\log(\frac{1}{\mu})}}{\sqrt{d}}\right)$ we have that

$$\Pr_{(x,y) \sim \mathcal{D}} (x \in T) = O\left(\eta \cdot \sqrt{\log\left(\frac{1}{\mu}\right)}\right) \quad (10)$$

Also, by equation (8) and the assumption that $\eta \leq \frac{1}{2(\alpha_0+1)}$, we have that $0 \leq \theta \leq \frac{\pi}{2}$. For this regime, $\sin(\theta) \geq \frac{2\theta}{\pi}$. Hence, by equation (6) we have

$$\frac{\sin(\theta)}{2\gamma\sqrt{d}} \geq \frac{\theta}{\pi\gamma\sqrt{d}} \geq \frac{\mu\eta}{\gamma\sqrt{d}} = \Theta\left(\frac{\mu}{\sqrt{\log(1/\mu)}}\right) \quad (11)$$

By equations (10) and (11) we can choose $\beta = \frac{\mu}{4C\sqrt{\log(\frac{1}{\mu})}}$, where $C > 0$ is a universal constant that is large enough so that

$$\beta < \frac{\sin(\theta)}{2\gamma\sqrt{d}} \text{ and } 2\beta \cdot \Pr_{(x,y) \sim \mathcal{D}} (x \in T) \leq \frac{\mu\eta}{2} \quad (12)$$

By equation 12 and lemma 2.3 we can choose $r = \Theta\left(\frac{\log^2(\frac{1}{\beta})}{\beta^2}\right) = \Theta\left(\frac{\log^3(\frac{1}{\mu})}{\mu^2}\right)$ such that

$$\min_{P' \in \text{POL}_{r,d}} \|h_{w^*} - P'\|_{1,\mathcal{D}|_T} \leq \beta$$

in that case we have

$$\text{Err}_{\mathcal{D}|_T}(P) \leq \text{Err}_{\mathcal{D}|_T}(h_{w^*}) + \min_{P' \in \text{POL}_{r,d}} \|h_{w^*} - P'\|_{1,\mathcal{D}|_T} + \beta \leq \text{Err}_{\mathcal{D}|_T}(h_{w^*}) + 2\beta.$$

Hence,

$$\begin{aligned} \Pr_{(x,y) \sim \mathcal{D}} (x \in T) \cdot \text{Err}_{\mathcal{D}|_T}(P) &\leq \Pr_{(x,y) \sim \mathcal{D}} (x \in T) \cdot \text{Err}_{\mathcal{D}|_T}(h_{w^*}) + \Pr_{(x,y) \sim \mathcal{D}} (x \in T) \cdot 2\beta \\ &\leq \Pr_{(x,y) \sim \mathcal{D}} (x \in T) \cdot \text{Err}_{\mathcal{D}|_T}(h_{w^*}) + \frac{\mu\eta}{2} \\ &= \Pr_{(x,y) \sim \mathcal{D}} (h_{w^*}(x) \neq y \text{ and } |\langle w, x \rangle| \leq \gamma) + \frac{\mu\eta}{2} \end{aligned} \quad (13)$$

By equations (7), (9) and (13) we conclude that

$$\begin{aligned}
\text{Err}_{\mathcal{D}}(h) &\leq \frac{\mu\eta}{2} + \Pr_{(x,y) \sim \mathcal{D}}(h_{w^*}(x) \neq y \text{ and } |\langle w, x \rangle| > \gamma) \\
&\quad + \Pr_{(x,y) \sim \mathcal{D}}(h_{w^*}(x) \neq y \text{ and } |\langle w, x \rangle| \leq \gamma) + \frac{\mu\eta}{2} \\
&= \text{Err}_{\mathcal{D}}(h_{w^*}) + \mu\eta \leq (1 - \mu)\eta + \mu\eta = \eta.
\end{aligned}$$

□

3 Polynomial approximation of the sign function

In this section we will find ℓ_1 approximation of halfspaces. In particular, we will prove lemmas 2.3 and 2.2.

3.1 Approximation in “truncated L^∞ ”

Lemma 3.1 *Let $a, \gamma, \tau > 0$. There exist a polynomial p of degree $O\left(\frac{1}{\gamma} \cdot \log\left(\frac{1}{\tau}\right)\right)$ such that*

- For $x \in [-a, a]$, $|p(x)| < 1 + \tau$.
- For $x \in [-a, a] \setminus [-\gamma \cdot a, \gamma \cdot a]$, $|p(x) - \text{sign}(x)| < \tau$.

We will use the following lemma:

Lemma 3.2 *Let $\tau > 0$. There exist a polynomial p of degree $O\left(\log\left(\frac{1}{\tau}\right)\right)$ such that*

- For $x \in [-1.5, 1.5]$, $|p(x)| < 1 + \tau$.
- For $x \in [-1.5, 1.5] \setminus [-0.5, 0.5]$, $|p(x) - \text{sign}(x)| < \tau$.

Proof The proof is established by approximating the error function, $\text{erf}(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$ by a low degree polynomial. Let $\sigma = 2\sqrt{2 \log\left(\frac{4}{\sqrt{2\pi}\tau}\right)}$. We claim that for every $x > \frac{\sigma}{2}$ we have

$$|\text{erf}(x) - 1|, |\text{erf}(-x)| \leq \frac{\tau}{4}. \quad (14)$$

Because $0 \leq \text{erf}(x) \leq 1$ for all x , and since $\text{erf}(x) = 1 - \text{erf}(-x)$, it is enough to prove that $\text{erf}(x) \geq 1 - \frac{\tau}{4}$. Indeed, we have

$$\begin{aligned}
1 - \text{erf}(x) &= \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{t^2}{2}} dt \\
&\leq \frac{1}{\sqrt{2\pi}} \int_x^\infty t e^{-\frac{t^2}{2}} dt \\
&= \frac{1}{\sqrt{2\pi}} \left[-e^{-\frac{t^2}{2}} \right]_x^\infty \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\
&\leq \frac{1}{\sqrt{2\pi}} e^{-\frac{\sigma^2}{2}} = \frac{\tau}{4}.
\end{aligned}$$

Now, by the Taylor expansion of e^x we have

$$e^{-\frac{x^2}{2}} = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{n! 2^n}.$$

Integrating element-wise and using the fact that $\text{erf}(0) = \frac{1}{2}$, we have

$$\text{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{n! 2^n (2n+1)}.$$

Let r be the $2k$ 'th Taylor polynomial of erf for $k = \max\{\lceil 2(1.5\sigma)^2 e \rceil, \log_2(\frac{4}{\tau})\} = O(\log(\frac{1}{\tau}))$. We have, for $|x| \leq 1.5\sigma \leq \sqrt{\frac{k}{2e}}$

$$\begin{aligned} |r(x) - \text{erf}(x)| &\leq \frac{2}{\sqrt{\pi}} \sum_{n=k}^{\infty} \frac{|x|^{2n+1}}{n! (2n+1)} \\ &\leq \frac{2}{\sqrt{\pi}} \sum_{n=k}^{\infty} \frac{x^{2n}}{n!} \\ &\leq \frac{2}{\sqrt{\pi}} \sum_{n=k}^{\infty} \frac{x^{2n}}{\sqrt{2\pi} \left(\frac{n}{e}\right)^n} \\ &\leq \frac{\sqrt{2}}{\pi} \sum_{n=k}^{\infty} \left(\frac{x^2 e}{n}\right)^n \\ &\leq \frac{\sqrt{2}}{\pi} \sum_{n=k}^{\infty} \left(\frac{1}{2}\right)^n \\ &= \frac{\sqrt{2}}{\pi} \left(\frac{1}{2}\right)^{k-1} \leq \left(\frac{1}{2}\right)^k \leq \frac{\tau}{4} \end{aligned}$$

Here, the 4'th inequality follows from the well known fact that $n! \geq \sqrt{2\pi} \left(\frac{n}{e}\right)^n$. Finally, using the last inequality and equation (14), it is not hard to check that the polynomial $p(x) = 2r(\sigma x) - 1$ satisfies the required properties.

□

Proof (of lemma 3.1) By rescaling, we can assume w.l.o.g. that $a = 1$. Let $\phi : [-1, 1] \rightarrow \mathbb{R}$ be the function

$$\phi(x) = \begin{cases} \frac{1}{\gamma} x & |x| \leq \gamma \\ 1 & x \geq \gamma \\ -1 & x \leq -\gamma \end{cases}$$

By Jackson's Theorem, there is a polynomial $q : [-1, 1] \rightarrow \mathbb{R}$ of degree $\leq \left\lceil \frac{12}{\gamma} \right\rceil$ with $\|q - \phi\|_{\infty, [-1, 1]} \leq \frac{1}{2}$. Also, let r be the polynomial from Lemma 3.2. It is easy to check that, $p = r \circ q$ satisfies the requirement of the Lemma.

□

3.2 Approximations for short tailed distributions

Lemma 3.3 Let $\rho : \mathbb{R} \rightarrow \mathbb{R}_+$ a density function such that for some $\gamma, \sigma > 0$ we have

$$\forall x, \rho(x) \leq \frac{2}{\sigma} \text{ and } \forall |x| > 2\gamma, \rho(x) \leq \frac{2}{\sigma} \exp\left(-\frac{x^2}{32\sigma^2}\right)$$

Then, for every $0 < \tau \leq \frac{\sigma}{2\gamma}$ there is a polynomial of degree⁵ $O\left(\frac{\log^2(1/\tau)}{\tau^2}\right)$ such that

$$\int_{-\infty}^{\infty} |p(x) - \text{sign}(x)|\rho(x)dx \leq \tau$$

We will use the following fact.

Lemma 3.4 [5] Let $p : \mathbb{R} \rightarrow \mathbb{R}$ be a polynomial of degree $\leq r$ for which $|p(x)| \leq b$ in the interval $[-a, a]$. Then, for every $|x| \geq a$ we have $|p(x)| \leq b \cdot \left|\frac{2x}{a}\right|^r$.

Proof (of lemma 3.3) By lemma 3.1, there is a polynomial p of degree $O(r \log(1/\tau))$ such that

- For $x \in [-r\tau\sigma, r\tau\sigma]$, $|p(x)| < 2$.
- For $x \in [-r\tau\sigma, -\frac{\tau\sigma}{100}]$, $|p(x)| < \frac{\tau}{100}$.
- For $x \in [\frac{\tau\sigma}{100}, r\tau\sigma]$, $|p(x) - 1| < \frac{\tau}{100}$.

We have,

$$\begin{aligned} \int_{-\infty}^{\infty} |p(x) - \text{sign}(x)|\rho(x)dx &= \int_{|x| < \frac{\tau\sigma}{100}} |p(x) - \text{sign}(x)|\rho(x)dx + \int_{\frac{\tau\sigma}{100} \leq |x| \leq r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx \\ &\quad + \int_{|x| \geq r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx \\ &\leq \int_{|x| < \frac{\tau\sigma}{100}} \frac{6}{\sigma} dx + \int_{\frac{\tau\sigma}{100} \leq |x| \leq r\tau\sigma} \frac{\tau}{100} \rho(x)dx \\ &\quad + \int_{|x| \geq r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx \\ &\leq \frac{\tau}{2} + \int_{|x| \geq r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx \end{aligned}$$

It remains to bound $\int_{|x| \geq r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx$. We will choose $r \geq \frac{1}{\tau^2}$, and therefore we will have $r\tau\sigma \geq \frac{\sigma}{\tau} \geq 2\gamma$. Hence, by lemma 3.4 we have

$$\begin{aligned} \int_{|x| \geq r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx &\leq \int_{|x| \geq r\tau\sigma} 3 \left(\frac{2x}{r\tau\sigma}\right)^r \frac{2}{\sigma} e^{-\frac{x^2}{32\sigma^2}} dx \\ &\leq 12 \int_{r\tau\sigma}^{\infty} \left(\frac{2x}{r\tau\sigma}\right)^r \frac{1}{\sigma} e^{-\frac{x^2}{32\sigma^2}} dx \\ &= 12 \int_{r\tau}^{\infty} \left(\frac{2y}{r\tau}\right)^r e^{-\frac{y^2}{32}} dy \\ &\leq 12 \int_{r\tau}^{\infty} \left(\left(\frac{2y}{r\tau}\right)^r e^{-\frac{y^2}{64}}\right) e^{-\frac{y^2}{64}} dy \end{aligned}$$

⁵The constant in the big-O notation is universal.

Now, it is possible to choose $r = \Theta\left(\frac{\log(1/\tau)}{\tau^2}\right)$ such that for all $y > r\tau$ we have $\left(\frac{2y}{r\tau}\right)^r \cdot e^{-\frac{y^2}{64}} \leq 1$. For such r , the last expression is bounded by $12 \int_{\omega(\frac{1}{\tau})}^{\infty} e^{-\frac{y^2}{64}} dy = o(\tau)$.

□

3.3 Approximation on a biased strip: proof of lemma 2.3

In this section we will find a low degree approximation of halfspaces w.r.t. to the distribution from step 3 of our PTAS. Namely, we will prove lemma 2.3. Let $\rho_{d,\gamma,\theta} : [-1, 1] \rightarrow \mathbb{R}_+$ be the projection on w^* of the uniform distribution on $T_{d,\gamma}(w)$. By equation (2), it is enough to find τ -approximation of the sign function in ℓ_1 , w.r.t. $\rho_{d,\gamma,\theta}$. Namely, it is enough to prove:

Lemma 3.5 *There is a univariate polynomial p of degree $r = O\left(\frac{\log^2(1/\tau)}{\tau}\right)$ such that*

$$\int_{-1}^1 |\text{sign}(x) - p(x)| \rho_{d,\gamma,\theta}(x) dx \leq \tau.$$

Lemma 3.5 follows immediately from lemma 3.3 with $\sigma = \frac{\sin(\theta)}{\sqrt{d}}$, the assumptions that $\gamma < \frac{1}{2}$ and $\tau < \frac{\sin(\theta)}{2\gamma\sqrt{d}}$, and the following bound:

Lemma 3.6

$$\begin{aligned} \forall z, \rho_{d,\gamma,\theta}(z) &\leq \frac{\sqrt{d}}{\sin(\theta)\sqrt{1-\gamma^2}} \\ \forall |z| \geq \gamma, \rho_{d,\gamma,\theta}(z) &\leq \frac{\sqrt{d}}{\sin(\theta)\sqrt{1-\gamma^2}} \exp\left(-(d-1)\frac{(|z|-\gamma)^2}{4\sin^2(\theta)}\right) \end{aligned}$$

To prove lemma 3.6, we will use an explicit formula for $\rho_{d,\gamma,\theta}$. It will be convenient to introduce some notation. Let $\rho_{d,r} : \mathbb{R} \rightarrow \mathbb{R}$ be the density function of the random variable that is the inner product of a fixed unit vector in S^{d-1} and a uniform vector in $r \cdot S^{d-1}$. Clearly,

$$\rho_{d,r}(x) = \frac{1}{r} \cdot \rho_{d,1}\left(\frac{x}{r}\right) \quad (15)$$

We will use the following well known inequality

$$\rho_d(x) \leq \sqrt{d} \exp\left(-\frac{x^2 d}{4}\right) \quad (16)$$

Lemma 3.7 *Let A be the probability of $T_{d,\gamma}(w)$ according to the uniform distribution. We have*

$$\rho_{d,\gamma,\theta}(z) = \frac{1}{A} \int_{-\gamma \cos(\theta)}^{\gamma \cos(\theta)} \rho_{d,\cos(\theta)}(u) \cdot \rho_{d-1,\sqrt{\sin^2(\theta)-\tan^2(\theta)u^2}}(z-u) du$$

Proof Let x be a uniform vector in the strip $T_{d,\gamma}(w)$, and let $y = \langle w^*, x \rangle$. We note that $\rho_{d,\gamma,\theta}$ is the density of y . We write

$$x = \alpha \cdot w + z$$

where $\langle w, z \rangle = 0$. For $(w^*)^\perp = w^* - \langle w^*, w \rangle w$ we have,

$$\begin{aligned} y = \langle w^*, x \rangle &= \alpha \cdot \langle w^*, w \rangle + \langle w^*, z \rangle \\ &= \alpha \cdot \cos(\theta) + \langle (w^*)^\perp, z \rangle \end{aligned}$$

We note that the density function of the distribution of $\alpha \cdot \cos(\theta)$ is given by

$$\tau(u) = \begin{cases} \frac{1}{A} \rho_{d, \cos(\theta)}(u) & |u| \leq \gamma \cdot \cos(\theta) \\ 0 & |u| > \gamma \cdot \cos(\theta) \end{cases}$$

Now, given α , z is a uniform vector of norm $\sqrt{1 - \alpha^2}$ in the orthogonal complement of w , and $(w^*)^\perp$ is a vector of norm $\sin(\theta)$ in that space. It follows that the density function of $\langle (w^*)^\perp, z \rangle$ given that $\alpha \cdot \cos(\theta) = u$ is $\rho_{d-1, \sin(\theta) \cdot \sqrt{1 - \frac{u^2}{\cos^2(\theta)}}} = \rho_{d-1, \sqrt{\sin^2(\theta) - \tan^2(\theta)u^2}}$. It therefore follows that

$$\rho_{d, \gamma, \theta}(z) = \frac{1}{A} \int_{-\gamma \cos(\theta)}^{\gamma \cos(\theta)} \rho_{d, \cos(\theta)}(u) \cdot \rho_{d-1, \sqrt{\sin^2(\theta) - \tan^2(\theta)u^2}}(z - u) du$$

□

We are now ready to prove lemma 3.6.

Proof (of lemma 3.6) Let A be the probability of the strip $T_{d, \gamma}(w)$ according to the uniform distribution on the sphere. We have, using equations (15) and (16),

$$\begin{aligned} \rho_{d, \gamma, \theta}(z) &= \frac{1}{A} \int_{-\gamma \cos(\theta)}^{\gamma \cos(\theta)} \rho_{d, \cos(\theta)}(u) \cdot \rho_{d-1, \sqrt{\sin^2(\theta) - \tan^2(\theta)u^2}}(z - u) du \\ &\leq \frac{1}{A} \int_{-\gamma \cos(\theta)}^{\gamma \cos(\theta)} \rho_{d, \cos(\theta)}(u) \cdot \rho_{d-1, \sqrt{\sin^2(\theta) - \tan^2(\theta)u^2}}(0) du \\ &\leq \frac{\sqrt{d-1}}{\sin(\theta) \sqrt{1 - \gamma^2}} \end{aligned}$$

Similarly, for $|z| > \gamma$,

$$\begin{aligned} \rho_{d, \gamma, \theta}(z) &= \frac{1}{A} \int_{-\gamma \cos(\theta)}^{\gamma \cos(\theta)} \rho_{d, \cos(\theta)}(u) \cdot \rho_{d-1, \sqrt{\sin^2(\theta) - \tan^2(\theta)u^2}}(z - u) du \\ &\leq \frac{1}{A} \int_{-\gamma \cos(\theta)}^{\gamma \cos(\theta)} \rho_{d, \cos(\theta)}(u) \cdot \rho_{d-1, \sqrt{\sin^2(\theta) - \tan^2(\theta)u^2}}(|z| - \gamma) du \\ &\leq \frac{1}{A \sin(\theta) \sqrt{1 - \gamma^2}} \int_{-\gamma \cos(\theta)}^{\gamma \cos(\theta)} \rho_{d, \cos(\theta)}(u) \cdot \rho_{d-1, 1} \left(\frac{|z| - \gamma}{\sqrt{\sin^2(\theta) - \tan^2(\theta)u^2}} \right) du \\ &\leq \frac{1}{A \sin(\theta) \sqrt{1 - \gamma^2}} \int_{-\gamma \cos(\theta)}^{\gamma \cos(\theta)} \rho_{d, \cos(\theta)}(u) \cdot \rho_{d-1, 1} \left(\frac{|z| - \gamma}{\sin(\theta)} \right) du \\ &= \frac{1}{\sin(\theta) \sqrt{1 - \gamma^2}} \rho_{d-1, 1} \left(\frac{|z| - \gamma}{\sin(\theta)} \right) \\ &\leq \frac{\sqrt{d}}{\sin(\theta) \sqrt{1 - \gamma^2}} \exp \left(-(d-1) \frac{(|z| - \gamma)^2}{4 \sin^2(\theta)} \right) \end{aligned}$$

□

Proof (of lemma 2.2) By equation (2), it is enough to show that there is a univariate polynomial p of degree $r = O\left(\frac{\log^2(1/\tau)}{\tau^2}\right)$ such that

$$\int_{-1}^1 |p(x) - \text{sign}(x)| \rho_{d, 1}(x) dx \leq \tau.$$

This, however, follows immediately from lemma 3.3 and equation (16).

□

Acknowledgements: Amit Daniely is a recipient of the Google Europe Fellowship in Learning Theory, and this research is supported in part by this Google Fellowship. The author thanks Pranjal Awasthi, Adam Klivans, Nati Linial, and Shai Shalev-Shwartz for valuable discussions and comments.

References

- [1] Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 724–733. IEEE, 1993.
- [2] Pranjal Awasthi, Maria-Florina Balcan, and Phil Long. The power of localization for efficiently learning linear separators with noise. In *STOC*, 2014.
- [3] P.L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.
- [4] S. Ben-David, D. Loker, N. Srebro, and K. Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. In *ICML*, 2012.
- [5] I. Ben-Eliezer, S. Lovett, and A. Yadin. Polynomial threshold functions: Structure, approximation and pseudorandomness. *Unpublished manuscript*, 2009.
- [6] A. Birnbaum and S. Shalev-Shwartz. Learning halfspaces with the zero-one loss: Time-accuracy tradeoffs. In *NIPS*, 2012.
- [7] Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Arxiv preprint arXiv:1505.05800 v1*, 2015.
- [8] Amit Daniely and Shai Shalev-Shwartz. Complexity theoretic limitations on learning dnf’s. In *Arxiv preprint arXiv:1404.3378 v1*, 2014.
- [9] Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. From average case complexity to improper learning complexity. In *STOC*, 2014.
- [10] Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. The complexity of learning halfspaces using generalized linear methods. In *COLT*, 2014.
- [11] Philip J Davis. *Interpolation and approximation*. Courier Dover Publications, 1975.
- [12] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A Servedio, and Emanuele Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010.
- [13] Ilias Diakonikolas, Daniel M Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 11–20. IEEE, 2010.
- [14] V. Feldman, P. Gopalan, S. Khot, and A.K. Ponnuswami. New results for learning noisy parities and halfspaces. In *In Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 2006.

- [15] V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. In *Proceedings of the 47th Foundations of Computer Science (FOCS)*, 2006.
- [16] A. Kalai, A.R. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Foundations of Computer Science (FOCS)*, 2005.
- [17] Michael Kearns and Ming Li. Learning in the presence of malicious errors. pages 267–280, May 1988. *SIAM Journal on Computing*.
- [18] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.
- [19] Adam Klivans and Pravesh Kothari. Embedding hard learning problems into gaussian space. In *RANDOM*, 2014.
- [20] Adam R Klivans, Ryan O’Donnell, and Rocco Servedio. Learning intersections and thresholds of halfspaces. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, pages 177–186. IEEE, 2002.
- [21] A.R. Klivans, P.M. Long, and R.A. Servedio. Learning halfspaces with malicious noise. *The Journal of Machine Learning Research*, 10:2715–2740, 2009.
- [22] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. In *FOCS*, pages 574–579, October 1989.
- [23] P.M. Long and R.A. Servedio. Learning large-margin halfspaces with more malicious noise. In *NIPS*, 2011.
- [24] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52:55–66, 2010.
- [25] S. Shalev-Shwartz, O. Shamir, and K. Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM Journal on Computing*, 40:1623–1646, 2011.
- [26] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.